Nucleic Acids Research, Vol. 19, No. 17 4731–4738

Medium reiteration frequency repetitive sequences in the human genome

David J.Kaplan, Jerzy Jurka[1], Joseph F.Solus and Craig H.Duncan*
Center for Molecular Biology, Wayne State University, Detroit, MI and [1]Linus Pauling Institute of Science and Medicine, 440 Page Mill Road, Palo Alto, CA 94306, USA

Received April 24, 1991; Revised and Accepted August 7, 1991 EMBL accession nos X59017–X59026 (incl.)

ABSTRACT

Fourteen novel medium reiteration frequency (MER) families were found, in the human genome, by using two different methods. Repetition frequencies per haploid human genome were estimated for each of these families as well as for six previously described MER DNA families. By these measurements, the families were found to contain variable numbers of elements, ranging from 200 to 10,000 copies per haploid human genome.

INTRODUCTION

The human genome, like those of other higher eukaryotes, contains a substantial amount of interspersed repetitive DNA sequence. Besides the three largest families, the Alu family, the L1 family, and the THE repeats, there are a number of minor families, referred to as medium reiteration frequency repeats (MER).

Medium frequency repetitive sequences were first observed as sequences which were present in naturally occurring SV40 variants (1,2). With the recent surge of structural information about the human genome, more examples of MER families have been uncovered. Two quite different approaches have proven useful for identifying MER families. One method was a systematic computer analysis of the entire human DNA sequence database to detect repetitive elements which do not belong to known families (3). The second method was a sequence library construction technique which selected regions of human DNA lying between closely spaced Alu repeats (4). In this report, we use both approaches to discover fourteen novel MER families. We also provide measurements of the repetition frequency of each family. There are certainly more families which have yet escaped detection.

EXPERIMENTAL METHODS

Homology searches by computer assisted analysis of the human DNA database

The GenBank DNA sequence database (Release 65) was subjected to a rapid search using the DASHER2 program as previously described (3).

Isolation of novel MER families from sequence libraries

The Alu fragment library. Human genomic DNA (250 μg) was digested to completion with the restriction enzyme *Alu*I. The resulting fragments were fractionated on a 6% polyacrylamide gel and fragments in the 500–1000 bp region were isolated. 50 ng of this size fractionated DNA was ligated to 500 ng of *Sma*I cleaved M13mp19 RF DNA (5). The vector was phosphatased before use. The ligated DNA was mixed with competent JM109 bacteria (Stratagene Inc.) and 20,000 resulting transformants were plated at a density of 1,650 plaques per 10 cm petri dish. Duplicate filter replicates were prepared and were incubated separately with either of two radioactive oligonucleotide probes. One of these oligonucleotides was the 25 mer, GTGGCTCA[C/T][A/G]CCTGTAATCCCAGCA. This sequence is from the 5′ consensus sequence for Alu repeats (bases # 12–36 from Fig. 1, ref 6). The other oligonucleotide was the 33 mer GG[C/T]TGCAGTGAGC[C/T][A/G][T/A]GAT[C/T][A/G][C/T][A/G]CCA[C/T]TGCACT. This sequence was from the 3′ region of the Alu family consensus sequence (bases # 218–250 from Fig. 1, ref 6). Phage which hybridized to both probes were replated at lower density and subjected to a repeat screening with the same probes. Single stranded template DNA was extracted from 131 phage isolates in 2 ml cultures.

The PCR library. One μg of genomic human placental DNA was mixed with 20 μg each of two oligonucleotides. One oligonucleotide was the 31 mer GGGTCGACAGTGAGCCGAGATCGCGCCACTG, where the underlined region is an outward facing primer at the 3′ boundary of the Alu family consensus sequence (bases 224–246 from Fig. 1, ref 6). The second oligonucleotide was the 28 mer GGGGATCCTGGGATTACAGGCGTGAGCC, where the underlined region is an outward facing primer at the 5′ boundary of the Alu family consensus sequence (bases 33–14 from Fig. 1, ref 6). The reaction mixture was adjusted to a volume of 200μl, containing each dNTP at a final concentration of 300μM, polymerase reaction buffer, and 40 units of *Thermus aquaticus* DNA polymerase (Amersham). The polymerase chain reaction (7) was performed for 25 cycles of 94° for 3 minutes, followed by 50° for 5 minutes, followed by 72° for 5 minutes. The reaction was then extracted twice with an equal volume of phenol, once with

* To whom correspondence should be addressed at: Room 4118, MCHT, 2727 2nd Avenue, Detroit, MI 48201, USA

an equal volume of $CHCl_3$, and ethanol precipitated. The DNA was then dissolved in 50 μl of restriction enzyme buffer, followed by the addition of 25 units of *Sal*I and 50 units of *Bam*H1. After a 4 hour incubation at 37°, the reaction mixture was heated at 65° for 5 minutes and DNA fragments were purified by two cycles of preparative electrophoresis on a 7% polyacrylamide gel. DNA fragments in the 300−1000 bp region were isolated. 40 ng of this size fractionated DNA was ligated to 100 ng of *Bam*H1, *Sal*I cleaved M13mp19 RF DNA (from which the 12 bp *Sal*I-*Bam*H1 insert had been removed by S-300 gel filtration [Pharmacia]). The ligated DNA was mixed with competent JM109 bacteria (Stratagene Inc.) and 2000 resulting transformants were plated at a density of 150 plaques per 10 cm petri dish on NZY plates containing β-galactosidase indicator dye. Duplicate filter replicates were prepared and were incubated separately with either of two radioactive probes. One probe was a double stranded 69 bp sequence taken from the internal region of an Alu repeat (CACTTTGGGAGGCCAAGGCAGGTGGATCACCTCAA-GTCAGGAGTTCAAGGCCAGCCTGACCAACATGGA). The second probe was a 5 kb fragment containing an L1 sequence from the region 5′ to the human γ-gamma globin gene (8). 431 phage plaques were selected by the criteria of not reacting with the dye indicator and not hybridizing to either of the radioactive probes. Single stranded template DNA was extracted from small scale (3 ml) cultures of these phage.

DNA, sequencing, amplification, hybridization, and sequence analysis

The single stranded templates were sequenced by the dideoxy termination method (8) as modified (9) using T7 DNA Polymerase (10). 400−600 bp of sequence information was obtained from each template. Oligonucleotides were synthesized by Research Genetics Inc (Huntsville AL) and labeled with polynucleotide kinase and ^{32}P labeled γ-ATP. Double stranded DNA was labeled by the random primer method (11). Hybridizations were performed at 60° in 6×SSC, 20 mM $NaPO_4$, 10% dextran sulfate, 4×Denhart's solution, 0.1% SDS, and 100 μg/ml denatured salmon sperm DNA for 16 hrs. Washing was done twice for 5 minutes in 2×SSC, 0.2% SDS at room temperature, followed by two washes for 15 minutes in 0.2×SSC, 0.1% SDS at 45°. Autoradiography was 6 hrs to overnight with an intensifying screen at −70°. MER probes were constructed by synthesizing oligonucleotide primers suitable for Polymerase Chain Reaction amplification (7) of the M13 sequence templates. The PCR products were then subcloned in Bluescript vectors (Stratagene Inc.). The sequence data were analysed on a Macintosh II microcomputer using MacVector 3.5 DNA analysis software from International Biotechnologies Inc. Southern blot hybridizations were performed according to Southern (12) as modified by Thomas (13).

Reiteration Frequency of DNA probes

Reiteration frequencies of the probes were estimated by a *plaque hybridization assay*. Purified double stranded DNA fragments used as probes were provided by Lagan Inc., Detroit MI. These DNA probes were labeled with ^{32}P by the random primer method and purified by centrifugal chromatography on Sephadex G-25 (15). The probes were then hybridized *in situ* to 25,000−50,000 recombinant lambda bacteriophage plaques from a genomic human DNA library which had been immobilized on a 150 mm circular nitrocellulose filter (14). The reiteration frequency of a particular probe was then estimated by the formula, $\{(\#\text{ of positive plaques binding probe}) \times (2.5 \times 10^6)\} \div \{(\text{total }\#\text{ of plaques on the filter}) \times 15\}$. In this formula the 2.5×10^6 represents the size of the human genome in kb, while 15 is the average size (in kb) of the human DNA inserted in the bacteriophage. As reflected by the numbers in Table I, this formula gives a statistical estimate, which may be in error by as much as 50%.

RESULTS

Sequence Analysis of novel MER families

A total of 562 M13 templates were subjected to sequence analysis. 131 of these templates came from the Alu fragment library and 431 of the templates came from the PCR library. In both cases, the selection procedures were designed to identify DNA sequences 200−700 bp in length which are flanked by Alu repeats.

The Alu fragment method was based on the fact that the bulk of Alu family elements in the human genome contain a conserved site for the restriction enzyme *Alu*I (16). When human genomic DNA was cleaved with this enzyme a subset of the resulting fragments possessed the 3′ portion of an Alu family repeat at one end, the 5′ portion of an Alu family repeat at the other end, with an internal region of non-Alu family DNA in between. The *Alu*I restriction fragments were cloned and the subset of DNA fragments described above was identified by the use of two oligonucleotide hybridization probes. One was specific for the 5′ end of an Alu family repeat, and the other was specific for the 3′ end of an Alu family repeat.

The PCR method used two oligonucleotide primers which faced outward from the 5′ and 3′ regions of a consensus Alu family repeat sequence. When these primers were incubated with human genomic DNA under PCR conditions, the regions between Alu repeats were selectively amplified. These amplified fragments were then cloned.

Both these methods suffered limitations which precluded the isolation of certain sequences or included undesired sequences

Table I. MER sequences.

MER#	Source	# in GenBank	Repetition Frequency	# in Alu frag lib	# in PCR lib	Hyb to rod DNA	Hyb to bov DNA
1	DASHER2: ref 3	4	4000-8000	-	-	—	+
2	DASHER2: ref 3	4	1500-3000	-	2	—	+
3	DASHER2: ref 3	-	nd	-	-	nd	nd
4	DASHER2: ref 3	3	1000-2000	-	-	—	—
5	DASHER2: ref 3	2	1000-2000	-	1	—	+
6	DASHER2: ref 3	2	500-1000	-	-	—	—
7	Alu frag lib: ref 4	3	500-1000	1	-	—	—
8	Alu frag lib	2	400-800	1	-	—	+
9	DASHER2	2	700-1500	-	-	—	—
10	DASHER2: ref 26	4	4000-8000	-	1	—	—
11	DASHER2	3	1500-3000	-	-	—	—
12	PCR lib	4	2500-5000	-	4	—	+
13	DASHER2	5	1500-3000	-	1	—	—
14	PCR lib	2	1500-3000	-	1	—	—
15	PCR lib	5	700-1500	-	2	—	+
16	PCR lib	2	300-600	-	1	—	+
17	PCR lib	1	1500-3000	-	1	—	—
18	Alu frag lib: ref 35	5	5000-10000	1	1	—	+
19	Alu frag lib	1	2500-5000	1	-	—	—
20	Alu frag lib	1	200-400	1	-	—	+
21	Alu frag lib	1	500-1000	1	-	—	+
TOTALS		56	31300-62800	6	14		

Characteristics of 21 MER sequences are presented, including the source by which they were discovered, the number of their occurrences in GenBank 65, their repetition frequencies by the plaque hybridization assay, the numbers of their occurrences in both the Alu fragment and PCR libraries, and their hybridization signals to rodent and bovine chromosomal DNAs. DASHER2 is the computer program for rapid similarity searches. The repetition frequencies are given in copies per haploid human genome. MER3 was a loosely homologous sequence family detected by DASHER2 analysis, but a probe made from this family did not detect any repetitive sequences by the plaque hybridization assay.

in the selection of clones. In some ways, the two methods were complementary in this regard. The main weakness of the PCR library stemmed from the fact that many Alu family elements exist as directly repeated tandem multimers. Any such arrangement formed a short (30−40 bp) sequence which was an ideal candidate to be amplified during the polymerase chain reaction using the primers we employed. The polymerase chain reaction products were subjected to two cycles of preparative gel electrophoresis before cloning, to select inserts of greater than 200 bp. Despite this, only 220 of 431 clones contained inserts of greater than 150 bp. Only 178 of these clones contained inserts of greater than 100 bp of non-Alu family sequence.

The Alu fragment library was designed to contain pieces of Alu family repeats at the boundaries with non-Alu family DNA in between. However, 57 of the 131 clones isolated by the Alu fragment method contained otherwise intact Alu family repeats which did not possess the characteristic *Alu*I restriction enyzme cleavage site. In most cases, the sequence data obtained from such clones consisted of Alu family sequence with little or no flanking regions. A further weakness of this method was that the clones were designed to terminate at all *Alu*I restriction sites.

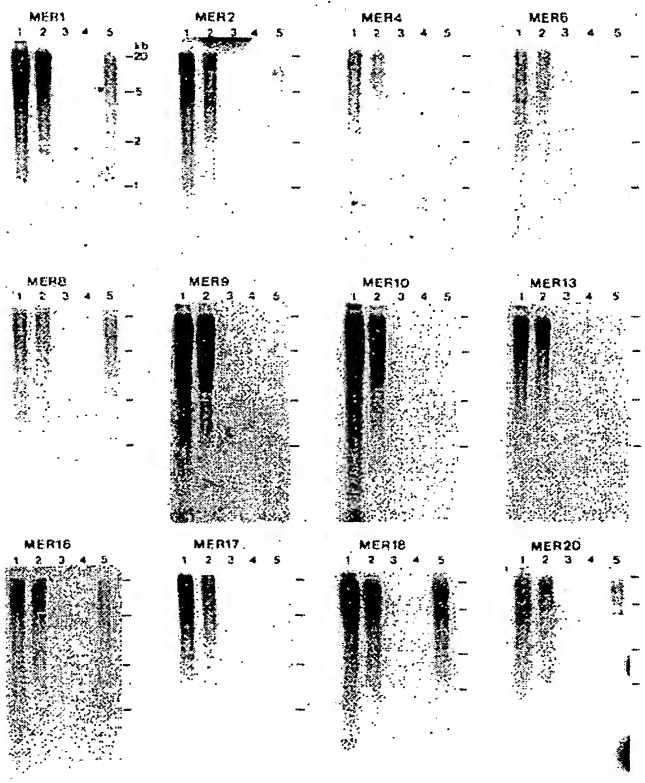


Figure 1. Hybridization of mammalian chromosomal DNAs with MER DNA probes. Five μg samples of restriction endonuclease digested chromosomal DNAs were electrophoresed in a 0.7% agarose gel and blotted onto nylon membranes. The membranes were incubated with radiolabelled DNA probes described in Table III or Ref. 3. Shown are autoradiographs of the blots. The lanes are: 1, Human DNA digested with *Eco*R1; 2, Rhesus monkey DNA digested with *Eco*R1; 3, Mouse DNA digested with *Eco*R1; 4,Chinese hamster DNA digested with *Eco*R1; 5, Bovine DNA digested with *Eco*R1. Positions of molecular weight markers are shown by dashes at the right of each autoradiograph.

Thus any region of the genome spanning an *Alu*I restriction site was excluded from this library.

Sequence data was obtained from all 562 clones selected from both libraries. This sequence data was then aligned against a datafile containing other known primate repetitive elements, including L1 sequences (17), satellite sequences (18), Xba repeats (19), O repeats (20), THE repeats (21), simple polydinucleotides and homopolymer runs greater than 12 bp. All regions homologous to these elements were removed. At this point the sequence datafile contained 58,272 bp, of which 22,285 bp were derived from the Alu fragment library and 35,987 bp were derived from the PCR library. This sequence was divided into blocks of 4−5 kb and aligned with the primate sequence database.

Table II. Locations of MERs in known chromosomal loci.

Locus Name	Alu sequences	MER #	length in bp	kb per MER
HUMAGG 727-981	←	1	4669	
HUMAIGRE 295-410	–	10	412	
HUMAPOAI 6055-6395	→	1	8967	
HUMAPOA4C 1-145	–	2	3606	
HUMATP1A2 4820-4980	–	16	26669	
HUMB2M2 1863-2030	–	7	2872	
HUMBCR22I 1090-1220	→	12	3023	
HUMC21DLA 1395-1560	–	18	2717	
HUMC21DLA 1600-1800	–	15	"	
HUMCRYGBC 4720-4877	→	4	22776	5.7
HUMCRYGBC 9067-9180	××	17	"	
HUMCRYGBC 9524-9921	○	9	"	
HUMCRYGBC 21020 21145	–	13	"	
HUMCYAR01 1010-1120	–	21	1297	
HUMCYP345 2880-3063	←	12	3064	
HUMERYA 2165-2478	–	2	3075	
HUMFIXG 6669-7025	→	2	38060	7.6
HUMFIXG 8000-8258	→	7	"	
HUMFIXG 18220-18400	–	13	"	
HUMFIXG 21700-22145	–	9	"	
HUMFIXG 24021-24170	→	5	"	
HUMFOLA2 1-264	–	6	596	
HUMGSTP1A 1-571	–	11	2670	
HUMHBB 14860-15020	–	13	73327	36.7
HUMHBB 58220-58340	–	13	"	
HUMHLASBA 8810-9025	–	10	14647	
HUMHMG14A 8044-8248	»	7	8883	
HUMHPRTB 7812-8071	××	18	56737	14.2
HUMHPRTB 37050-37250	→	12	"	
HUMHPRTB 54772-54975	←←	15	"	
HUMHPRTB 54960-55145	←←	18	"	
HUMHXAMUT 1-150	←	14	352	
HUMIGCD3 283-500	–	10	762	
HUMIL2RBA 170-320	–	14	1096	
HUMINSRD 1060 1200	»	13	7341	
HUMINT2 11440 11560	–	15	11809	
HUMLMWO61 390-540	→	16	1482	
HUMMBP1A 273-400	–	15	3032	
HUMNGFB 8399-8550	–	5	11595	
HUMP45C17 53-790	–	11	8550	
HUMPADP 45-240	–	18	3805	
HUMPAIA 1-511	–	1	15868	
HUMRASSK2 530-750	–	19	2444	
HUMRPB17A 3245-3500	–	20	4030	
HUMSEXREPB 350-560	–	18	917	
HUMSIGMG3 403-1194	–	11	3724	
HUMSIGMG3 2000-2150	–	15	"	
HUMTPPB 8913-8989	→	8	13866	
HUMTPA 40-578	→	1	36594	6.2
HUMTPA 5307-6297	ins	4	"	
HUMTPA 17474-17877	→	12	"	
HUMTPA 18392-18851	ins	2	"	
HUMTPA 22931-23031	→	8	"	
HUMTPA 24498-25144	→	4	"	
HUMTPA 34152-35153	ins	6	"	
HUMTPOO4 451-673	–	10	4019	
Alu fragment library			22285	3.7
PCR library			35987	2.6

The 56 MERs with known genomic locations are listed. All sequences use the GenBank numbering system. The column *Alu sequences* refers to the presence or absence of Alu family repeats with 500 bp 5′ or 3′ of the given genomic location and the arrows refer to the orientation of the Alu repeats. For example → means that there is no Alu sequence 5′ to the genomic location and that there is an Alu 3′ to the genomic location which is oriented in the 5′ to 3′ direction. The letters 'ins' refer to an insertion of an Alu element within the MER element. MER # is the MER designation of each repetitive element. Length in bp refers to the total length of the given genomic sequence in the database. kb per MER is obtained by dividing the length in kb of each loci by the numbers of MERs identified within selected loci.

Twenty seven sequences which showed matches to the primate database (greater than 150 on initial scoring or 300 on optimized scoring) were investigated further. Of these twenty seven sequences, nine were eliminated because they were members of previously described MER families. Double stranded DNA probes (100–300 bp) were constructed from the eighteen remaining sequences. Eleven of these probes identified repeat sequence families by the plaque hybridization assay (Materials & Methods). Probes were also constructed for ten additional MER families discovered by DASHER2 analysis of the GenBank 65 database. The estimated reiteration frequencies of all these probes are given (Table I).

The screening processes used here relied on the occurrence of homologous regions in the human sequence database. At least one region of homology was required with the sequence library methods and at least two regions of homology were required with the DASHER2 method. It was conceivable that some of the sequences in the Alu fragment or PCR libraries could be MER elements which are not represented in the human sequence database. To test this hypothesis probes were constructed for sixteen sequences from the libraries which did not have a significant match with the human primate database. None of these probes identified repeat sequence families in the plaque hybridization assay.

Hybridization of MERs to chromosomal DNAs

Genomic blot hybridizations were performed in order to evaluate the distribution of MER repeats in genomic DNA. Double stranded DNA probes for all MER families were radiolabelled and hybridized to *Eco*R1 digests of primate, bovine, Chinese hamster, and mouse chromosomal DNAs. The resulting autoradiographs (twelve representative films are shown, Fig. 1) showed hybridization of these probes to disperse sets of restriction fragments from human and rhesus monkey chromosomal DNAs. Similar results have been reported for another MER1 probe (22). Some of the patterns showed what appeared to be discrete length bands within a background of heterogeneous length fragments. Further investigation is required to explain the origin of these discrete bands. None of the probes showed hybridization to mouse or Chinese hamster chromosomal DNAs. The bovine chromosomal DNA showed hybridization signals for ten of the probes. These signals were always less intense than those of from an equivalent amount of primate DNA, but contrasted to no detectable signal in the rodents' chromosomal DNAs.

DISCUSSION

Isolation of MERs from the sequence libraries

The sequence library methods were designed to select for novel MERs on the basis of their proximity to Alu family repeats. Of the 56 known genomic MER locations, 26 lie within 500 bp of an Alu repeat (Table II). The Alu fragment and PCR libraries were constructed in such a way to select for regions of the genome adjacent to Alu repeats. Both of these sequence libraries were clearly enriched for MERs. The 59 kb of sequence analysed contained 20 examples of MERs or one repeat for every 2.9 kb. This is in contrast to known genomic loci containing MERs. Several of the long known gene sequences contain multiple examples of MERs, but even the most richly endowed sequence (HUMTPA) has fewer MERs per unit length (5.2 kb per MER) than either the PCR or Alu fragment libraries (Table II).

Description of MER families

Sequence alignments are presented for fourteen families, MER8 to MER21 (Table III). MER1 to MER6 (3) and MER7 (4) sequence alignments have been previously presented.

MER8. This sequence was present in the Alu fragment library. It is 60% homologous to a 160 bp region in the tissue plasminogen activator gene (23), and has a 70 bp homology to the tissue factor β gene (24).

MER9. This sequence was found by DASHER2 analysis of GenBank, as a 88% homology over 270 bp. The probe for this family was made from a sequence within intron e of the human gene for blood clotting factor IX (25).

MER10. This sequence was first noted as being repetitive by Lawrance et al. (26). It appears four times in the human sequence database, and different family members share about 80% homology. MER10 has also been studied by Mermer et al. (27) who referred to this family as the *Mst*II repeats. The probe for this family was constructed from an intergenic region in the human HLA locus (26).

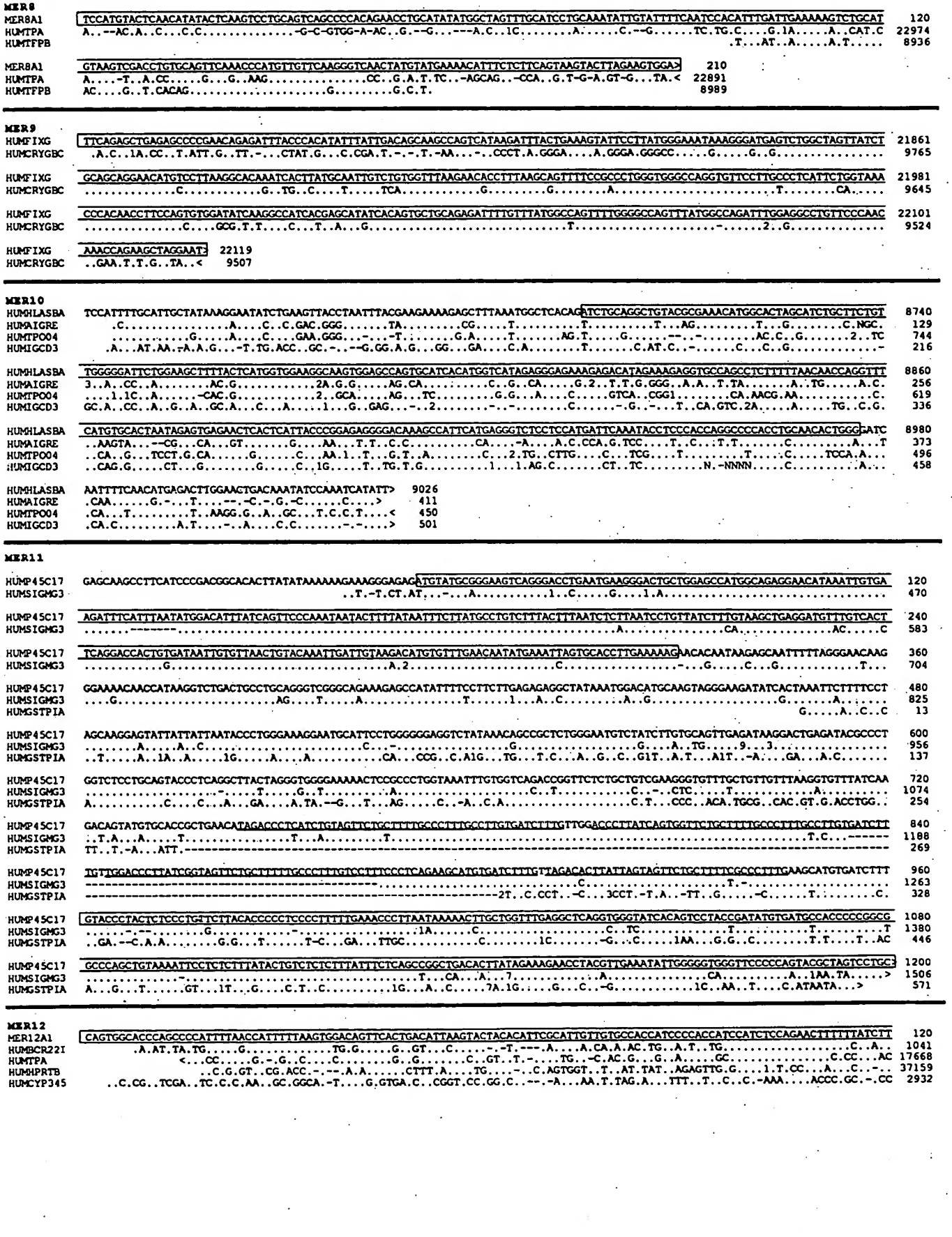
MER11. This sequence was found by DASHER2 analysis of GenBank. This is the longest MER found, showing matches over 1100 bp in three different genes. MER11 sequences exhibit length differences. This is due to the presence of variable numbers of a 50 bp subrepeat, which is present four times in the HUMP45C17 sequence (marked by underlines in Table I), three times in the HUMSIGMG3 sequence and once in the HUMGSTPIA sequence. Akahori et al. (28) noticed this 50 bp subrepeat in the HUMSIGMG3 sequence. Two probes for this family were constructed, one from the 5′ region and another from the 3′ region of the MER11 repeat in the human gene for cytochrome P450-C17 (29). Both probes indicated similar repeat frequencies in the plaque hybridization assay.

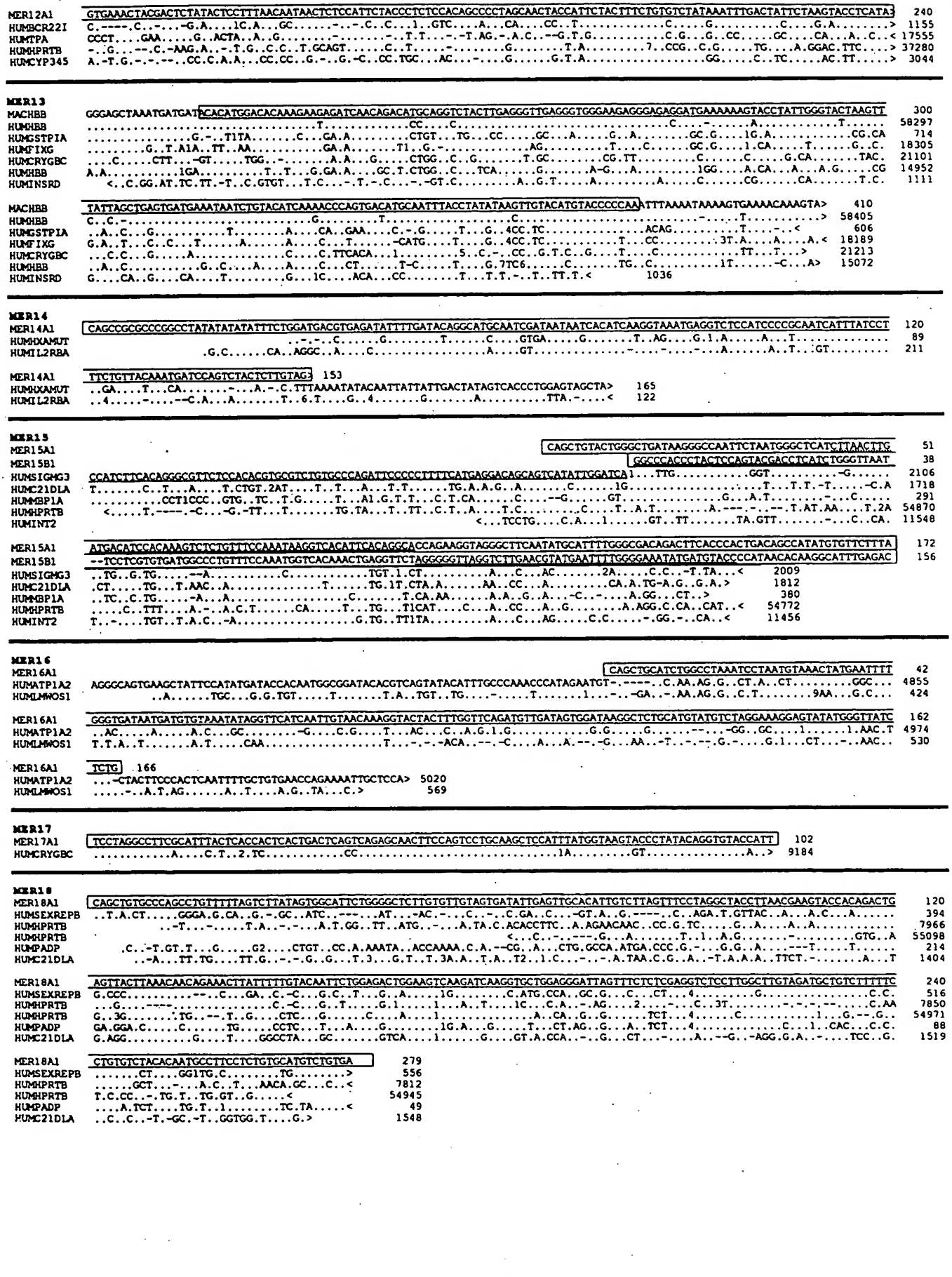
MER12. This sequence was found in the PCR library. There are four appearances of this family in the primate database, each sharing 60–70% homology with the others.

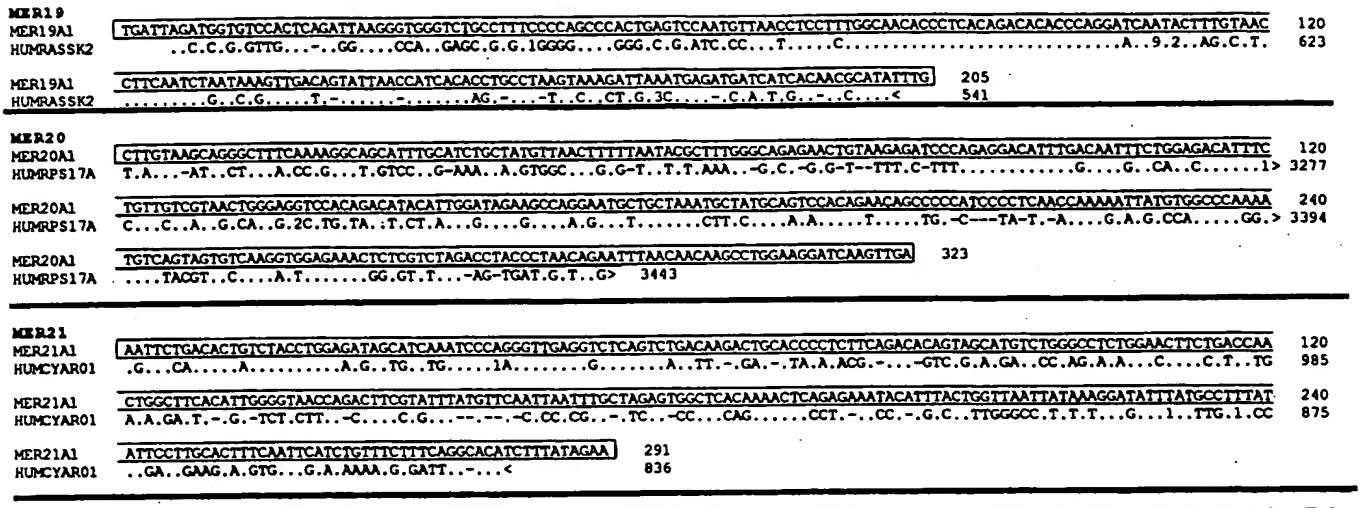
MER13. This sequence was found by computer analysis of the database. A probe was constructed from an intergenic region of the macaque β globin gene region (30). The probe shared 93% homology with its orthologous counterpart in the human β globin region and 70–75% homology to the other members of this family. The MER13 sequence may be part of a variant L1 sequence. It was considered to be the 3′ portion an L1 repeat in the human gene for blood clotting factor IX (25). It lies adjacent to, but was not considered part of, an L1H repeat in the human β-globin locus (31). MER13 is also adjacent to an L1H repeat in the human crystallin gene locus (32) and in 5′ flanking sequence of the human glutathione S-transferase pi gene (33). However MER13 was not associated with L1 repeats in its other two known genomic locations.

MER14. This sequence was found in the PCR library. It shares 75–80% homology with two sequences in the primate database. At one location MER14 flanks the 5′ side of a deletion in a mutant gene for the alpha chain of β-hexosaminidase A (34). The 3′ side of this deletion is within an Alu family repeat.

Table III. Alignment of MER8–21 with their locations at sequenced loci.




MER12A1
HUMBCR22I
HUMTPA
HUMHPRTB
HUMCYP345
MER13
MACHBB
HUMHBB
HUMGSTPIA
HUMFIXG
HUMCRYGBC
HUMHBB
HUMINSRD
MER14
MER14A1
HUMHXAMUT
HUMIL2RBA
MER15
MER15A1
MER15B1
HUMSIGMG3
HUMC21DLA
HUMMBP1A
HUMHPRTB
HUMINT2
MER16
MER16A1
HUMATP1A2
HUMLMWOS1
MER17
MER17A1
HUMCRYGBC
MER18
MER18A1
HUMSEXREPB
HUMHPRTB
HUMPADP
HUMC21DLA



The best match alignments from the MacVector programs are shown. Matches are indicated by periods, while mismatches are indicated by the letters G,A,T,C or dashes. Insertions are indicated by numbers showing their lengths. The boxed sequences indicate the regions used as probes in hybridization experiments. Nomenclature was adopted to allow future identification of additional members or subfamilies of MER repeats. For example MER8A1 indicates example number 1 of the A subfamily of MER8. The new sequences presented here have been assigned the following EMBL accession numbers: [MER12A1; X59017], [MER14A1; X59018], [MER15A1; X59019], [MER16A1; X59020], [MER15B1; X59021], [MER17A1; X59022], [MER19A1; X59023], [MER18A1; X59024], [MER21A1; X59025], [MER20A1; X59026]. Arrows (<or>) indicate the 5′ to 3′ directions of the sequence numberings. For MER15 the best consensus sequence is a mosaic of sequences and is indicated by dotted underlines. For MER11, underlines denote the 50 bp subrepeat described in DISCUSSION.

MER15. This sequence was noted by Akahori et al. (28) as being part of a repeat unit in the human immunoglobulin locus. Two sequences, MER15A1 and MER15B1, were found in the PCR library which shared homology with this repeat unit, which was called the MER15 family. Although MER15A1 and MER15B1 share only short patches of homology with each other, together they form a mosaic consensus sequence. Searches of the database identified five sequences which shared 75−80% homology with this consensus. MER15 sequences are found directly adjacent to MER18 sequences at two locations (HUMHPRTB 54970 and HUMC21DLA 1560). However, the two sequences are not juxtaposed at other genomic locations and hybridization experiments indicate different repetition frequencies, so MER15 and MER18 are considered to be distinct MER families.

MER16. This sequence was found in the PCR library. It shares 70% homology with two human sequences from the database.

MER17. This sequence was found in the Alu fragment library. Although short, it shares 87% homology with an intergenic sequence in the human crystallin gene locus (32).

MER18. This sequence was found in the PCR library. It is a member of a repeat family first described by Fisher et al. (35). This family is rather loosely conserved (60−65% homology) and is spread throughout the chromosomes. A more tightly conserved (93% homology) subfamily is embedded in tandem duplications on the human sex chromosomes (35). MER18 sequences are found directly adjacent to MER15 sequences at two locations (HUMHPRTB 54970 and HUMC21DLA 1560), but is still considered as a distinct MER family.

MER19. This sequence was found in the Alu fragment library. It contains a 100 bp region which is 85% homologous to a region 5′ to the human SK2 c-Ha-ras-1 oncogene (36).

MER20. This sequence was found in the Alu fragment library. Hybridization experiments indicate a low (200−400) copy number per haploid genome. One of these hybridization targets may be a 60% homology to the first intron of the human gene for ribosomal protein S17 (37).

MER21. This sequence was found in the Alu fragment library. As for MER20, hybridization experiments indicate a repetitive sequence, but database searches revealed only a 60% homology to a 100 bp sequence in the 5′ flank of the human aromatase cytochrome P450 gene (38). Such a homology may not be sufficient to have been detected in our hybridization experiments.

Undiscovered MER families

The 21 known MER families account for 30,000−60,000 repetitive elements in the human genome (Table I). There are surely more MER families remaining to be discovered. Their number can be estimated in several ways. The simplest way is to assume that the sequence libraries constructed in this study select MERs at random. Then, out of all the MERs in the libraries, the proportion which are members of already known MER families will correspond to the proportion of the MERs in the human genome which have already been assigned to known MER families. Specifically, the sequence libraries contained 20 matches to the GenBank database. Of these 20, 9 were members of known MER families. This implies that 9/20 or 45% of the MERs are in known families and that 55% remain to be discovered. This implies the existence of 70,000−140,000 MERs in the human genome.

Another way to estimate the abundance of MERs is to extrapolate from the abundance of known MERs within sequenced gene loci. This estimate would have to be a lower bound, because it neglects the not yet discovered MERs which may lie within these loci. Nonetheless, some rough figures can be assembled. Five gene loci (demarcated by boxes in Table II)

contain 227494 bp of DNA and 22 MERs. This extrapolates to 242,000 MERs in the entire human genome.

Although both of these estimates are crude, they give some notion of the number and complexity of MER families in the human genome. More work is required in many areas before the nature and function of MER families will be understood. Even at this early stage several interesting features stand out. One is the apparent clustering of MER repeats at certain genetic loci (Table II). This clustering may imply that MER families influence gene expression, although one experimental test of this idea failed to support it (22). Another interesting feature is the diversity in levels of homology between the different MER families, ranging from >90% (MER9, MER11) to <70% (MER12, MER18). This implies that some of the MER families are of recent evolutionary origin while others are ancient. Some of the MER families may be old enough to be relics of Alu-type repeats from the early times of the mammalian radiation. This latter interpretation is supported by the observation that ten of the MER probes showed hybridization signals with bovine chromosomal DNAs (Fig. 1).

Medium reiteration frequency, or MER repeats, are significant additions to the repertoire of interspersed repetitive DNA which has heretofore been described in the human genome. As the DNA databases grow in size, more MER families will be discovered. These families merit further study both because of their intrinsic importance as well as for their use as mapping sites for developing a physical map of the human genome.

ACKNOWLEDGEMENTS

The authors wish to thank Gayathri Murali for technical support. We are also grateful to the following colleagues who provided plasmids: S.Degen for the tissue plasminogen activator gene, K.Kurachi for the blood clotting factor IX gene, S.Lawrance for the HLA gene, and W.Miller/S.Brentano for the cytochrome P450-C17 gene. This research was supported by the State of Michigan Research Excellence Fund.

REFERENCES

1. Wakamiya, T., McCutchan, T., Rosenberg, M., and Singer, M. (1979) *J. Biol. Chem.*, **254**, 3584–3591.
2. Sheflin, L., Celeste, A., and Woodworth-Gutai, M. (1983) *J. Biol. Chem.*, **258**, 14315–14321.
3. Jurka, J (1990) *Nucleic Acids Res.*, **18**, 137–141.
4. Kaplan, D.J., and Duncan, C.H. (1990) *Nucleic Acids Res.*, **18**, 192.
5. Yanisch-Perron, C., Vieira, J., and Messing, J. (1985) *Gene*, **33**, 103–119.
6. Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S., and Matsubara, K. (1987) *Gene*, **53**, 1–10.
7. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H.A. (1988) *Science*, **239**, 487–491.
8. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
9. Brunner, A.M., Schimenti, J.C., and Duncan, C.H. (1986) *Biochemistry*, **25**, 5028–5035.
10. Tabor, S., and Richardson, C.C. DNA (1987) *Proc. Natl. Acad. Sci. U.S.A.*, **84**, 4767–4771.
11. Feinberg, A.P., and Vogelstein, B. (1983) *Anal. Biochem.*, **132**, 6–13.
12. Southern, E.M. (1975) *J. Mol. Biol.*, **98**, 503–517.
13. Thomas, P.S. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 5201–5205.
14. Benton, W.D., and Davis, R.W. (1977) *Science*, **196**, 180–182.
15. Neal W.M. and Florini, J.R. (1973) *Anal. Biochem.*, **55**, 328–330.
16. Houck, C.M., Rinehart, F.P. and Schmid, C.W.(1979) *J. Mol. Biol.*, **132**, 289–306.
17. Martin, S.L., Voliva, C.F., Burton, F.H., Edgell, M.H., and Hutchison, III, C.A. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 2308–2312.
18. Waye, J.S., England, S.B., and Willard, H.F. (1987) *Mol. Cell. Biol.*, **7**, 349–356.
19. Safrany, G. and Hidvegi E.J. (1989) *Nucleic Acids Res.*, **17**, 3013–3022.
20. Sun, L., Paulson, K.E., Schmid, C.W., Kadyk, L., and Leinwand, L. (1984) *Nucleic Acids Res.*, **12**, 2669–2690.
21. Paulson, K.E., Deka, N., Schmid, C.W., Misra, R., Schindler, C.W., Rush, M.G., Kadyk, L. and Leinwand, L. (1985) *Nature (London)*, **316**, 359–361.
22. Bosma, P.J., Kooistra, T., Siemieniak, D.R., and Slightom, J.L. (1991) *Gene* **100**, 261–266.
23. Degen, S. J. F., Rajput, B., and Reich, E. (1986) *J. Biol. Chem.*, **261**, 6972–6985.
24. Mackman, N., Morrissey, J.H., Fowler, B., and Edgington, T.S. (1989) *Biochemistry*, **28**, 1755–1762.
25. Kurachi, K., Davie, E.W., Strydom, D.J., Riordan, J.F., and Vallee, B.L. (1985) *Biochemistry*, **24**, 5494–5499.
26. Lawrance, S.K., Das, H.K., Pan, J., and Weissman, S.M. (1985) *Nucleic Acids Res.*, **13**, 7515–7528.
27. Mermer, B., Colb, M., and Krontiris, T.G. (1987) *Proc. Natl. Acad. Sci. U.S.A.* **84**, 3320–3324.
28. Akahori, Y., Handa, H., Imai, K. Abe, M., Kameyama, K., Hibiya, M., Yasui, H., Okamura, K., Naito, M., Matsuoka, H. and Kurosawa, Y. (1988) *Nucleic Acids Res.*, **16**, 9497–9511.
29. Picado-Leonard, J., and Miller, W.L. (1987) *DNA*, **6**, 439–448.
30. Savatier, P., Trabuchet, G., Chebloune, Y., Faure, C., Verdier, G., and Nigon, V.M. (1987) *J. Mol. Evol.*, **24**, 297–308.
31. Rogan, P.K., Pan J., and Weissman, S.M. (1987) *Mol. Biol. Evol.*, **4**, 327–342.
32. Den Dunnen, J.T., Van Neck, J.W., Cremers, F.P.M., Lubsen, N.H., and Schoenmakers, J.G.G. (1989) *Gene*, **78**, 201–213.
33. Morrow, C.S., Goldsmith, M.E. and Cowan, K.H. (1990) *Gene*, **88**, 215–225.
34. Myerowitz, R. and Hogikyan, N.D. (1987) *J. Biol. Chem.*, **262**, 15396–15399.
35. Fisher, E. M. C., Alitalo, T., Luoh, S-W., De la Chapelle, A., and Page, D.C. (1990) *Genomics*, **7**, 625–628.
36. Sekiya, T., Tokunaga, A., and Fushimi, M. (1985) *Jpn. J. Cancer Res.*, **76**, 787–791.
37. Chen, I.T. and Roufa, D.J. (1988) *Gene*, **70**, 107–116.
38. Means, G.D., Mahendroo, M.S., Corbin, C.J., Mathis, J.M., Powell, F.E., Mendelson, C.R., and Simpson, E.R. (1989) *J. Biol. Chem.*, **264**, 19385–19391.